UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Scalable Factorization Model to Discover Implicit and Explicit Similarities Across Domains

by

Duc Minh Quan Do

A Thesis Submitted for the Degree of

Doctor of Philosophy

Sydney, Australia

2018

UNIVERSITY OF TECHNOLOGY SYDNEY

SCHOOL OF SOFTWARE

The undersigned hereby certify that they have read this thesis entitled **"Scalable Factorization Model to Discover Implicit and Explicit Similarities Across Domains"** by **Duc Minh Quan Do** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Date:

Principal Supervisor: _____

Dr. Wei Liu

Certificate of Original Authorship

I, Duc Minh Quan Do declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Software, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) scholarship.

Date: 15/09/2018

Signature of Author: Production Note:
Signature removed prior to publication.

Acknowledgements

I am especially indebted to Dr. Wei Liu, who have provided continuous support, advice and invaluable comments to pursue my research goals. As my principal supervisor, he has guided me more than I could ever give him credit for here. Many thanks are also due to my co-supervisor, Dr. Fang Chen, for many useful discussions with her.

I am grateful to all I have had the pleasure to discuss with. Each of the members of my Candidature Assessment Committee has provided me a great deal of professional feedback about scientific research. This work would not have been possible without the financial support of the Commonwealth Scientific and Industrial Research Organisation Scholarship (formerly National ICT Australia Scholarship) and the UTS - International Research Scholarship (IRS).

Nobody has been more important to me in the pursuit of this thesis than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue and wherever I am. They are the ultimate role models. Most importantly, I am grateful to my loving and supportive wife, Yen, and my wonderful daughter, Ellen, for constant inspiration, patience, and faith.

For Ellen

My love for you will last forever.

Contents

List of Figures

List of Tables

List of Publications

Below is the list of journal and conference papers associated with my Ph.D. research:

1. **Quan Do**, Wei Liu, Fan Jin and Dacheng Tao, "Unveiling Hidden Implicit Similarities for Cross-Domain Recommendation," *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (Under review).
2. **Quan Do** and Wei Liu, "Scalable Multimodal Factorization for Learning from Very Big Data," *in Multimodal Analytics for Next-Generation Big Data Technologies and Applications*, Springer (To appear).
3. **Quan Do**, Wei Liu and Fang Chen, "Discovering both Explicit and Implicit Similarities for Cross-Domain Recommendation," *in Proceedings of the 2017 Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 618-630, May 23-26, 2017.
4. **Quan Do** and Wei Liu, "ASTen: an Accurate and Scalable Approach to Coupled Tensor Factorization," *in Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 99-106, Jul. 24-29, 2016.

Abbreviation

ALS Alternating Least Squares. 14

CBT Code Book Transfer. 6, 18

CF Collaborative Filtering. x, 1, 11

CLFM Cluster-Level Latent Factor Model. 6, 8, 18

CMF Collective Matrix Factorization. 2, 8, 13, 17

CMTF Coupled Matrix Tensor Factorization. 5, 6, 8, 11, 13–17

GD Gradient Descent. 14

GPU Graphics Processing Unit. 15

MF Matrix Factorization. 11, 12

NCG Nonlinear Conjugate Gradient. 14

NSW New South Wales. 1, 4

RMSE Root Means Squared Error. 6

TF Tensor Factorization. 11, 12, 14, 15

Nomenclature and Notation

A rating matrix from n users for m items is denoted by a boldface capital, e.g., $\mathbf{X}$. Each row of the matrix is a vector of a user's ratings for all items while each column is a vector of ratings from all users for a specific item. Vectors are denoted by boldface lowercases, i.e., $\mathbf{u}$. A boldface capital and lowercase with indices in their subscript are respectively used for an entry of a matrix and a vector. Table 1 lists all other symbols we thoroughly use in this thesis.

Table 1 : Symbols and their descriptions

Symbol	**Description**
$\mathbf{X}^{(i)}$	Rating matrix from i-th dataset
$\mathbf{U}^{(i)}$	The first dimension factor of $\mathbf{X}^{(i)}$
$\mathbf{V}^{(0)}$	Common parts of the coupled factors
$\mathbf{V}^{(i)}$	Domain-specific parts of the coupled factor of $\mathbf{X}^{(i)}$
$\mathbf{S}^{(i)}$	Weighting factor of $\mathbf{X}^{(i)}$
$\mathbf{A}^{\mathrm{T}}$	Transpose of $\mathbf{A}$
$\mathbf{A}^{\dagger}$	Moore-Penrose pseudo inverse of $\mathbf{A}$
$\mathbf{I}$	The identity matrix
$\|\mathbf{A}\|$	Frobenius norm
n, m, p	Dimension length
c	Number of common clusters in coupled factors
r	Rank of decomposition
$\Omega_{\mathbf{X}}$	Number of observations in $\mathbf{X}$
$\frac{\partial}{\partial \mathbf{x}}$	Partial derivative with respect to $\mathbf{x}$
$\mathcal{L}$	Loss function
λ	Regularization parameter
$\times$	Multiplication
x, $\mathbf{x}$, $\mathbf{X}$, $\boldsymbol{\mathcal{X}}$	A scalar, a vector, a matrix and a tensor
N	Mode of a tensor
M	Number of machine
K	Number of tensor
T	Number of iteration
$I_1 \times I_2 \times \cdots \times I_N$	Dimension of N-mode tensor $\boldsymbol{\mathcal{X}}$
$\|\Omega\|$, $\boldsymbol{\mathcal{X}}_{i_1,i_2,\ldots,i_N}$	Observed data size of $\boldsymbol{\mathcal{X}}$ and its entries
$\boldsymbol{\mathcal{X}}^{(n)}$	Mode n^{th} of $\boldsymbol{\mathcal{X}}$
$\boldsymbol{\mathcal{X}}^{(n)}_{i_n}$	Slice i_n of $\boldsymbol{\mathcal{X}}^{(n)}$ - all entries $\boldsymbol{\mathcal{X}}^{(n)}_{*,\ldots,*,i_n,*,\ldots,*}$
$\mathbf{U}^{(n)}$	n^{th} mode factor of $\boldsymbol{\mathcal{X}}$
$\mathbf{u}^{(n)}_{i_n}$	$i_{n^{\text{th}}}$ row of factor $\mathbf{U}^{(n)}$
$\mathbf{V}^{(2)}$	2^{nd} mode factor of $\mathbf{Y}$
$\mathbf{v}^{(2)}_{j_2}$	$j_{2^{\text{th}}}$ row of factor $\mathbf{V}^{(2)}$ - all entries $\mathbf{V}^{(2)}_{*,j_2}$
$\mathbf{U}1$, $\mathbf{U}2$,...,$\mathbf{U}K$	Factors of tensor $\boldsymbol{\mathcal{X}}1$, $\boldsymbol{\mathcal{X}}1$, $\boldsymbol{\mathcal{X}}K$
$I_1 \times I_2 \times \cdots \times I_{N_K}$	Dimension of N_K-mode tensor $\boldsymbol{\mathcal{X}}K$
$\|\Omega\|_K$, $\boldsymbol{\mathcal{X}}K_{i_1,i_2,\ldots,i_{N_K}}$	Observed data size of $\boldsymbol{\mathcal{X}}K$ and its entries

Abstract

E-commerce businesses increasingly depend on recommendation systems to introduce personalized services and products to their target customers. Achieving accurate recommendations requires a sufficient understanding of user preferences and item characteristics. Given the current innovations on the Web, coupled datasets are abundantly available across domains. An analysis of these datasets can provide a broader knowledge to understand the underlying relationship between users and items. This thorough understanding results in more collaborative filtering power and leads to a higher recommendation accuracy.

However, how to effectively use this knowledge for recommendation is still a challenging problem. In this research, we propose to exploit both explicit and implicit similarities extracted from latent factors across domains with matrix tri-factorization. On the coupled dimensions, common parts of the coupled factors across domains are shared among them. At the same time, their domain-specific parts are preserved. We show that such a configuration of both common and domain-specific parts benefits cross-domain recommendations significantly. Moreover, on the non-coupled dimensions, the middle factor of the tri-factorization is proposed to use to match the closely related clusters across datasets and align the matched ones to transfer cross-domain implicit similarities, further improving the recommendation.

Furthermore, when dealing with data coupled from different sources, the scalability of the analytical method is another significant concern. We design a distributed factorization model that can scale up as the observed data across domains increases. Our data parallelism, based on Apache Spark, enables the model to have the smallest communication cost. Also, the model is equipped with an optimized solver that converges faster. We demonstrate that these key features stabilize our model's per-

formance when the data grows.

Validated on real-world datasets, our developed model outperforms the existing algorithms regarding recommendation accuracy and scalability. These empirical results illustrate the potential of our research in exploiting both explicit and implicit similarities across domains for improving recommendation performance.